Discrimination, Allocatory and Separatory, Linear Aspects*

by

Seymour Geisser**

University of Minnesota

Technical Report No. 264

*To be delivered at the "Advanced Seminar on Classification and Clustering" on May 5, 1976 at the University of Wisconsin.

**Research supported by a grant from the Army Research Office.

Discrimination, Allocatory and Separatory, Linear Aspects

by

Seymour Geisser*

University of Minnesota

1. Introduction

The classification of objects is one of the hoariest and consequently not the least primitive of scientific enterprises - certainly considerably removed from preciser mechanisms directed towards the explanation and prediction of natural and social phenomena. Briefly, it attempts to sort out in some sensible manner objects belonging to two or more labeled classes. When this involves a parsimonious and efficient criterion of choice based on related manifest attributes, we are in the realm of Discrimination.

An early recorded instance appears in the biblical book, Judges, XII, 5-6. A clan of Israelites from Gilead held the fords over the Jordan to prevent the defeated troops of Ephraim, another Israelite tribe, from crossing the river. The Ephraimites sharing race, language, customs and dress were apparently indistinguishable in all respects from the Gileadites. Seizing upon a dialectical variation as an efficient sorting device, the guards made those attempting the ford pronounce the word "Shibboleth". Upon hearing "Sibboleth" they were fairly certain of apprehending an Ephraimite.

It is quite likely that their errors of classification were no greater

---

*Supported in part by U.S. Army Research Office

than many of our current weather classifiers aided by a modern computer, who base forecasts of snow on a large number of precisely determined variables. This is of course a situation where the label has in fact not yet occurred but is predictive as opposed to the previous retrodictive case.

Often in the latter case the latent label of a new object can only be ascertained with certainty by prodigeous experimental effort which may even involve the destruction or alteration of the object rendering it useless for further inquiry. Other cases may require an inordinate amount of time and patience until the label eventually reveals itself. Hence the utilization of easily assessed related attributes may be of invaluable aid in a study if only for reasons of economics and prudence.

There is also a natural hierarchy in terms of how these problems can be organized. In the least informative situations, the number of classes as well as the labels are unknown, and it is hoped that clues to both these entities will be disclosed by some set of appropriate manifest attributes. Here the basic problem is determining the number of classes and of forming clusters. In more informative cases the number of classes or populations is known or specified. Further knowledge is often also presumed concerning certain aspects of the attribute distributions.

For the sake of clarity we set down the general problem as follows: There are populations (or patterns) π_j , $j=1,\ldots,r$, with r known or unknown and π_j possibly specified by moments or by a distribution function $F_j(\cdot|\theta_j)$, whose form may be known or unknown, and θ_j is the j^{th} set of known or unknown parameters. There may be certain relationships among the

π_j , as well as subpopulations π_{ji} . Further there are two sets of observations, the first denoted by X and the second by U , (either set may be empty). Each of the observations belonging to X is such that its population origin or label is known with certitude, but the labels of those belonging to U are not. These may have some prior probabilities attached to them before they are observed and one object of the endeavor is to determine their origin in some optimal manner. Here allocation is the goal. A second goal, which may be primary in certain studies, is basically descriptive (graphical, algebraic or some other qualitative form), and involves initially the disclosure of the manifest differential features of the patterns, populations or potential populations under scrutiny. The purposes of the first are action oriented, predictive or retrodictive while the latter is more in the realm of the speculative in terms of possibly throwing some light on scientific or social issues.

In the first case one attempts to derive some rule which optimally allocates new observations while in the second instance one tends to focus on functions (discriminants) which tend to maximally distinguish or separate the populations. An appropriate allocatory procedure requires prior probabilities of an observation belonging to one or another population or estimates thereof. Often they are not obtainable and one tacitly assumes that these prior probabilities are equal. In many cases this is tantamount to using a separatory function as an allocator and the two original distinct goals tend to fuse or become blurred. Allocatory optimality is basically definable only when stringent assumptions are met while in vague situations a separatory function may sometimes usefully serve as an allocator. Conversely, allocatory notions may also be used to define a separatory function.

Discrimination, in its modern guise, was founded by R. A. Fisher (1936). He derived those linear functions of the class of all linear functions that best separated populations (actually samples) in terms of maximizing a certain distance function depending on only the first two moments.

Since then, linear discriminants have played an important role in the theory. From other points of view it was also found that linear theory was preeminent in one of the most useful of distributions, the multivariate normal, Wald (1944), Welch (1939).

In this paper we shall present not only an exposition of linear discrimination but shall also attempt to give a coherent discussion of its twin goals - allocation and separation.

In the next few sections we review linearity in the multivariate normal case, discuss the extent to which linearity is optimal and indicate the actual use of linear discriminants. This is followed by a section in which the distributional assumptions are dropped and the thrust is on the separation of populations via linear functions. An incidental feature is that some of the basic results are derived algebraically in a manner which differs from customary derivations. The penultimate section is devoted to the application of sample reuse procedures to linear discriminants.

2. Multivariate Normal Case.

Suppose there are p-dimensional multivariate populations $\pi_1,\ldots,\pi_r$ with vector means $\mu_1,\ldots,\mu_r$ and common positive definite covariance matrix Σ. One is interested in allocating a new p-dimensional observation u to one of these various populations in some optimal fashion. Assuming u has prior probability q_i of belonging to π_i, $\sum_{i=1}^{r} q_i=1$, then the optimal method for multivariate normal populations with regard to total posterior probability of correct classification (PCC), c.f. Anderson (1958) is to allocate u to that π_i for which

$$w_i(p) = \log q_i - \tfrac{1}{2}D_i^2(p) , \quad i=1,\ldots,r \tag{2.1}$$

is a maximum where

$$D_i^2(p) = (u-\mu_i)'\Sigma^{-1}(u-\mu_i) , \tag{2.2}$$

the Mahalanobis distance. This is the solution which allocates u to that π_i which has maximum posterior probability since $w_i(p)$ is easily shown to be a monotone function of $P_r[\pi_i|u]$, the posterior probability that u is from π_i.

It is sometimes of interest to determine whether we can transform linearly the set of p variables into $k \le p$ variables and preserve the allocation in k dimensions. Let $y = Cu$, $\eta_i = C\mu_i$, $\Omega = C\Sigma C'$, for C, a $k \times p$ matrix of rank $k \le p$, and

$$D_i^2(k) = (y-\eta_i)'\Omega^{-1}(y-\eta_i) = (u-\mu_i)'C'(C\Sigma C')^{-1}C(u-\mu_i) , \tag{2.3}$$

the corresponding distance in k dimensions.

Assume $\beta = \sum_{i=1}^{r} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'$ where $\bar{\mu} = r^{-1} \sum_{i=1}^{r} \mu_i$ and β is of rank $r-v \leq p$, noting that when $\mu_1, \ldots, \mu_r$ are linearly independent $v=1$. Since β is a p.s.d. matrix there exists a Λ such that, $\beta = \Lambda\Lambda'$, where Λ is $p \times r-v$. If we let $C = P'\Lambda'\Sigma^{-1}$ where $k = r-v$ and P, the $r-v \times r-v$ orthogonal matrix such that $P'\Lambda'\Sigma^{-1}\Lambda P$ is $\text{diag}(\delta_1, \ldots, \delta_{r-v})$ where δ_j are the non-zero roots of $\Sigma^{-1}\beta$ in descending order, then

$$D_i^2(k) = (u-\mu_i)'\Sigma^{-1}\Lambda[\Lambda'\Sigma^{-1}\Lambda]^{-1}\Lambda'\Sigma^{-1}(u-\mu_i) . \tag{2.4}$$

Further by adding and subtracting $\bar{\mu}$ in $u-\mu_i$ and noting that $\mu_i - \bar{\mu}$ is in the vector space generated by $\Lambda\Lambda'$ it is easily shown that for all i

$$w_i(p) - w_i(k) = D_i^2(p) - D_i^2(k) = (u-\bar{\mu})'[\Sigma^{-1} - \Sigma^{-1}\Lambda[\Lambda'\Sigma^{-1}\Lambda]^{-1}\Lambda'\Sigma^{-1}](u-\bar{\mu}) \tag{2.5}$$

and hence is independent of i. Therefore allocation of y by means of the maximum $w_i(k)$ is equivalent to the original allocation of u, thus verifying that $C = P'\Lambda'\Sigma^{-1}$ is a solution that preserves the original allocation. The new set of coordinates y are referred to as the complete set of linear multiple discriminants and they contain all of the discriminatory power of the original set of coordinates. The set y is an orthogonal set and forms a basis for all other solutions $y^* = Ry$ where R is any real non-singular $k \times k$ matrix. On the other hand if $k < r-v$, the allocation by y, the transform of u, will not be the same as the allocation by u for all u, as can easily be verified.

The total probability that u will be correctly allocated by the procedure is

$$q(p) = \sum_{i=1}^{r} q_i \Pr[u \in R_i | \pi_i] \tag{2.6}$$

where R_i is the region given by those u satisfying $\underset{j}{\text{Max}}\, w_j(p) = w_i(p)$, and is a maximum with respect to all possible procedures. If $y = P'\Lambda'\Sigma^{-1}u$ then it is clear that

$$q(p) = q(r\text{-}v) = \sum_{i=1}^{r} q_i \Pr[y \in R_i^* | \pi_i] \tag{2.7}$$

where R_i^* is given by those y satisfying $\underset{j}{\text{Max}}\, w_j(r\text{-}v) = w_i(r\text{-}v)$. To simplify matters let $q_i = r^{-1}$ for all i so that

$$q(p) = q(r\text{-}v) = r^{-1} \sum_{i=1}^{r} \Pr[y \in R_i^* | \pi_i] \tag{2.8}$$

then R_i and R_i^* are given by u and y which minimize $D_i^2(p)$ and $D_i^2(r\text{-}v)$ respectively. When $k < r\text{-}v$ and we use the procedure, i.e., minimizing $D_i^2(k)$ of (2.3) where $y = Cu$, then $q(k) < q(r\text{-}v)$ by continuity arguments. On the other hand it might be conjectured that the best one can do with respect to maximizing $q(k)$ is to let $C = P'_{(k)}\Lambda'\Sigma^{-1}$ where $P_{(k)} = (P_1, \ldots P_k)$ is the matrix of the first k columns of P, i.e., P_i is the invariant vector associated with the i^{th} largest root of $\Lambda'\Sigma^{-1}\Lambda$, or equivalently ΛP_i is the invariant vector associated with the identical root of $\beta\Sigma^{-1}$. This conjecture is in general false whenever $r\text{-}v \geq 2$ if we wish to maximize $q(k)$, as a counterexample will show. But from another point of view, i.e., optimizing on separatory criteria which we shall discuss in Section 5, it can be best.

A further note of caution should be introduced to the effect that the PCC is only of value in assessing the discriminatory power of the manifest variables

at hand prior to the observation of u. Once a set of such variables is determined and a particular u observed, the only relevant factor is the posterior probability, when calculable, that u belongs to one or another of $\pi_1,\ldots,\pi_r$,

$$\Pr[\pi_j|u] = q_j f_j(u)/\sum_{i=1}^{r} q_i f_i(u) \tag{2.9}$$

where $f_j(\cdot)$ represents in general the probability function associated with π_j.

We shall now describe the aforementioned counterexample. Suppose we ask for a single linear combination that will maximize the PCC assuming the r populations all have equal prior probability r^{-1}, blurring the distinction between allocation and separation. Then $c'u$, under π_j is univariate normal with mean $c'\mu_j$ and variance $c'\Sigma c$. Then $z = c'u/\sqrt{c'\Sigma c}$ is under π_j, $N(\eta_j,1)$ where $\eta_j = c'\mu_j/\sqrt{c'\Sigma c}$. Hence we can calculate the maximal probability of correct classification for any c

$$\text{PCC} = 2r^{-1}\sum_{i=1}^{r-1}\Phi\left(\frac{\eta_{(i)}-\eta_{(i+1)}}{2}\right) + (2-r)r^{-1} \tag{2.10}$$

where Φ is the distribution function of a standardized normal variate and $\eta_{(i)}$ are the ordered values of η_i such that $\eta_{(1)} \geq \eta_{(2)} \geq \ldots \geq \eta_{(r)}$. Maximization of the PCC with respect to c is troublesome, but it can be shown that the c that maximizes PCC is not necessarily the vector associated with the largest root of $\beta\Sigma^{-1}$ as one might initially suspect. Such a suspicion of course would arise from the fact that this vector does maximize the variation amongst the η_i. While this variation is contributory, the PCC is also quite sensitive to the spacing amongst the $\eta_{(i)}$.

The following example demonstrates these facts: Let u be a 3×1 vector with means under π_1, π_2 and π_3, respectively

$$\mu_1 = (1,0,0)\ ,\ \mu_2' = (0,1,-1)\ ,\ \mu_3 = (-1,-1,1) \quad \text{and} \quad \Sigma = I\ .$$

Then $\beta \Sigma^{-1} = \beta$ and

$$\beta = \sum_{j=1}^{3} (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})' = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & -2 \\ -1 & -2 & 2 \end{pmatrix}$$

with characteristic roots $3 + \sqrt{3}$, 0, $3 - \sqrt{3}$. The normed vector associated with the largest root is $c_1' = (6 - 2\sqrt{3})^{-\frac{1}{2}} (\sqrt{3} - 1, 1, -1)$. Using $c_1'u$ we find that $(\eta_{(1)}, \eta_{(2)}, \eta_{(3)}) = (6 - 2\sqrt{3})^{-\frac{1}{2}} (2, \sqrt{3} - 1, -1 - \sqrt{3})$ and compute the PCC to be .67757. On the other hand the simple normed vector $c_o' = (0, \frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ yields $(\eta_{(1)}, \eta_{(2)}, \eta_{(3)}) = (\sqrt{2}, 0, -\sqrt{2})$, equally spaced, and results in a PCC of .68033, which is just a trifle larger than that attained by the vector associated with the maximum root of β. In actual fact the normed vector $c^* = (.173, .697, -.697)$ leads to $(\eta_{(1)}, \eta_{(2)}, \eta_{(3)}) = (1.394, .173, -1.567)$ and yields a PCC of .69139 which is the maximum attainable here for a single linear combination. It is well known that the dispersion, $\sum_{i=1}^{r} (\eta_i - \bar{\eta})^2$ attains its maximum, $3 + \sqrt{3} = 4.732$ in this case, when the vector associated with the largest root is utilized, while the same measure of dispersion for the vector c_o is 4 - considerably less, and for the vector c^* we obtain 4.42.

Another way of viewing this problem is to realize that we are basically maximizing two quite different functions of the ordered values of $\eta_j = c'\mu_j / \sqrt{c' \Sigma c}$, $j=1, \ldots, r$ with respect to the arbitrary vector c. One

function is given by (2.10) while the other is

$$\sum_{j=1}^{r} (\eta_{(j)} - \bar{\eta})^2 \ . \tag{2.11}$$

That the characteristic vector associated with the largest root maximizes (2.11) results from the fact that (2.11) is invariant with regard to the ordering of the η_j so that from the definition of η_j we obtain

$$\sum_{j=1}^{r} (\eta_{(j)} - \bar{\eta})^2 = \sum_{j=1}^{r} (\eta_j - \bar{\eta})^2 = \frac{c'\beta c}{c'\Sigma c} \ . \tag{2.12}$$

As is well known, the quantity on the right of (2.12) is maximized when c is set equal to the characteristic vector associated with the largest root of $\beta\Sigma^{-1}$. Hence there is really no reason to expect the same solution for both cases.

We note that when $r = 2$, the optimal allocatory procedure yields the single linear discriminant

$$U^* = (u - \tfrac{1}{2}(\mu_1 + \mu_2))' \Sigma^{-1}(\mu_1 - \mu_2) + \log \frac{q_1}{q_2} \tag{2.13}$$

such that $U^* > 0$ assigns u to π_1 and $U^* \leq 0$ assigns u to π_2. Insertion of the usual estimates for μ_1, μ_2 and Σ when they are unknown and estimable from data yields the plug-in rule

$$V^* = (u - \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)' \hat{\Sigma}^{-1}(\hat{\mu}_1 - \mu_2) + \log \frac{q_1}{q_2} \tag{2.14}$$

with $V^* > 0$ assigning u to π_1 and π_2 otherwise.

It will be shown, however, that from certain points of view, even in this most structured of cases, linear theory, strictly speaking, may be inappropriate though approximately correct and certainly convenient.

3. The Limits of Linear Theory - Allocatory Aspects

For the remainder of our discussion we shall restrict ourselves to the two population case; π_1 and π_2 with density function $f(\cdot|\theta_1)$ and $f(\cdot|\theta_2)$ respectively. Optimal allocation for a new observation u with regard to the PCC involves assigning

$$u \text{ to } \pi_1 \text{ if } \rho(\theta,u) = \frac{q_1 f(u|\theta_1)}{q_2 f(u|\theta_2)} > 1 \qquad (3.1)$$
$$u \text{ to } \pi_2 \text{ otherwise}$$

where $\theta = \theta_1 \cup \theta_2$ is the entire set of distinct parameters of the problem. Equivalently any monotonic increasing function of ρ, say $h(\rho)$ for every fixed u will also do, so that any $h(\rho)$ may be denoted as an allocatory population discriminant. "Linear" theory is then surely optimal whenever there exists an $h(\rho)$ which is linear in u, although there are other cases as well. The multivariate normal distribution with equal covariance matrices is an example of the logistic class which always yields a linear population discriminant because of the form of

$$\rho(\theta,u) = e^{\alpha_o + \alpha' u} \qquad (3.2)$$

where $\alpha' = (\alpha_1,\ldots,\alpha_p)$ and consequently $\log \rho(\theta,u)$ is linear in u.

J. A. Anderson [1973] points out that multivariate independent dichotomous variables as well as several other interesting cases also belong to the logistic class. In fact this type of linearity remains valid for a special case of the general exponential family where θ_i is the set of parameters (β_i,τ) and

$$f(u|\theta_i) = g(\beta_i,\tau)h(u,\tau)e^{\beta_i' u} \qquad (3.3)$$

where β_i is a p-dimensional vector and τ is a set of extraneous parameters. But there are also other possibilities for linearity, e.g., two multivariate "student" distributions that differ in their location but have the same covariance matrices and equal prior probabilities Geisser [1966]. Here the rule (3.1) is equivalent to a rule linear in u derived from a positive root of $\rho(\theta,u)$. The rules conform exactly even though the positive root of $\rho(\theta,u)$ is nonlinear. For a slightly wider class of which the above is a special case see Enis and Geisser [1974]. Exact linear theory is then only strictly appropriate for restricted sets of distributional assumptions though somewhat wider than the logistic family. However, it is generally hoped that it will give reasonably robust, if less than optimal, solutions to many other cases. There are situations, however, where it certainly should not be applied, e.g., where two normal populations have the same mean but differ in their covariance matrices. Here linear discriminants will be quite inappropriate. This model reflects to a degree the situation arising in discriminating between fraternal and nonfraternal twins, see e.g., Richter and Geisser [1960], Okamoto [1961], Geisser and Desu [1968], Desu and Geisser [1973], Geisser [1973a].

However, except for special situations as just described, it is usually assumed or piously hoped that linearity will be at least a not unreasonable first approximation. By this is implied that the rule (3.1) can be replaced by a rule linear in u without great loss. For a contrary view in taxonomy see Reyment [1973]. In the classical frequential paradigm often an estimate of α is plugged into (5.1) while α_o is resolved into its constituent sum $\log q_1 q_2^{-1} + \alpha_o^*$ with an estimate for α_o^* plugged in and $\log q_1 q_2^{-1}$ assumed

to be a particular value, often 0 for convenience resulting in a discriminatory blur. Sometimes $q_1 q_2^{-1}$ is derived from a model, Geisser [1973a], or estimated from previous data or from the data at hand, when the situation permits Geisser [1964]. Now when α_o and α are known any $h(\rho)$, of course, will do as well. However, depending on which $h(\rho)$ and what is used for its estimation when the parameter values are only estimable from data, the sample discriminant or rule for allocation will in general vary. One way around this is to use maximum likelihood or any other estimator which will preserve the invariance of the rule. For a discussion of some of these and related points see Geisser [1969,1970] and Desu and Geisser [1973]. To do otherwise reqires that the statistician decide on whether the rule is paramount or the estimation of a particular discriminatory function $h(\rho)$ is crucial. Of course for large samples the discrepancy may be quite neglible.

However, as was noted, the logistic model itself encompasses a variety of possible distributional assumptions. While presumably robust for its class when its parameters are estimated it is not expected to yield as efficient a procedure when compared to one that is based on the true member of the class. For a logistic and normal comparison see Efron [1975].

Another classical approach, Wald [1944], Anderson [1958, 141-2], is via the testing of hypotheses. Here one computes the likelihood ratio test of the hypothesis that the new observation belongs to either of the two populations under scrutiny. More specifically if X_i is the set of observations known to be from π_i, then

$$\lambda = \frac{\underset{\theta}{\text{Max}}\ f(X_1|\theta_1)\ f(X_2|\theta_2)\ f(u|\theta_1)}{\underset{\theta}{\text{Max}}\ f(X_1|\theta_1)\ f(X_2|\theta_2)\ f(u|\theta_2)} \tag{3.4}$$

and u is assigned to π_1 if $\lambda > \frac{q_2}{q_1}$, π_2 otherwise. Hence $q_1 q_2^{-1}\lambda$ may

be termed the likelihood ratio allocatory discriminant.

For the multivariate normal case with equal covariance matrices,

$$\lambda = \left[\frac{1 + N_2\nu^{-1}(N_2+1)^{-1}(u-\bar{x}_2)'S^{-1}(u-\bar{x}_2)}{1 + N_1\nu^{-1}(N_1+1)^{-1}(u-\bar{x}_1)'S^{-1}(u-\bar{x}_1)}\right]^{(\nu+3)/2} \tag{3.5}$$

where $\bar{x}_i$ is the sample mean of N_i independent observations represented by X_i and known to have originated from π_i and S is the usual unbiased estimate of Σ with $\nu = N_1 + N_2 - 2$ degrees of freedom, $\nu > p$.

It is interesting to note that it is no longer necessarily possible to recover a linear discriminant from this procedure except under the rather restrictive assumption that $q_1^{2/(\nu+3)}N_2(N_2+1)^{-1} = q_2^{2/(\nu+3)}N_1(N_1+1)^{-1}$. Of course satisfaction is guaranteed if both $q_1=q_2$ and $N_1=N_2$. Although this may be disconcerting, it is not surprising as the thrust here is essentially on a rule (or test) rather than on the estimation of a true underlying linear population discriminant. Although the likelihood ratio discriminant for this paradigm is equivalent to a rule based on a quadratic discriminant it approaches linearity for large N_1 and N_2 so that for large enough samples there will be virtually little difference between it and the "usual" plug-in estimate (rule)

$$V^* = [u - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]' S^{-1}(\bar{x}_1 - \bar{x}_2) + \log \frac{q_1}{q_2} \tag{3.6}$$

for the true population discriminant (rule)

$$U^* = [u - \frac{1}{2}(\mu_1+\mu_2)] \Sigma^{-1}(\mu_1 - \mu_2) + \log \frac{q_1}{q_2}. \tag{3.7}$$

The rule indicated by (3.5) was shown to be an admissible Bayes rule by Kiefer and Schwartz [1965] and also Das Gupta [1965] for this allocation problem. However, the proper prior distribution which is utilized to prove the admissibility is one that most Bayesians would consider grossly deficient

in that it depends on the sum of the sample sizes and only assigns non-zero density to functions of Σ, μ_1, μ_2 in a space restricted to p-dimensions whereas the set (Σ, μ_1, μ_2) ostensibly contains $(1/2)p(p+5)$ parameters. This does not say too much for the Bayes admissible character of the rule. Whether a proper prior can be obtained which does not have these drawbacks is an open question, but it is not likely.

Another Bayesian derivation given by Geisser [1964] uses the simple improper prior density

$$g(\Sigma^{-1}, \mu_1, \mu_2) \propto |\Sigma|^{\frac{P+1}{2}}, \tag{3.8}$$

and also results in a quadratic rule in general. Hence V^* is not recoverable for arbitrary values of N_1, N_2, p, q_1 and q_2, Geisser [1966], but it is recoverable except for an additive constant depending on a particular relationship existing among these values, Enis and Geisser [1974]. It is only fully recoverable for the special case $N_1=N_2$ and $q_1=q_2$. Hence on a strictly allocatory basis the linear discriminant V has not been found to be admissible.

A semi-Bayesian justification, Geisser [1967], based on the aforementioned improper prior, focuses on the Bayesian estimation of U^*, rather than on allocation. This approach yields for the posterior expectation of U^*

$$E(U^*|u) = V^* + \frac{p}{2}(N_2^{-1} - N_1^{-1}) \tag{3.9}$$

which all but recovers the linear rule (5.6) and completely so whenever $N_1=N_2$. Elaborations of the use of this method are presented by Enis and Geisser [1970]. Another Bayesian approach Enis and Geisser [1974], which stresses linearity also yields results close to the rule V^*. Here one determines that linear function which maximizes the PCC with respect to the predictive distribution of the observation to be classified. Here an allocatory notion is utilized as the separa-

tory criterion with discriminants restricted to a linear class. This attempts to "optimally" compromise the allocatory needs with the desirability of linearity.

In both normal and non-normal applications, V^* is often utilized although it is not clear whether this emanates from the fact that the normal population discriminant U^* is linear and both allocatory and separatory and V^* is a good estimate of U^*, or that V^* for $q_1 = q_2$ can be derived as the "best" separatory linear discriminant in a distribution free setting utilizing the sample, Fisher [1936]. Basically it appears that for many less sophisticated users of the technique it is both the simplicity of linearity combined with the authority of Fisher that is compelling. At any rate, there seems to be a bias in applications (as well as theory) for focusing on linear discriminants rather than a quest for overall optimal allocation irrespective of the goal. One has only to peruse the discriminatory literature to observe that almost all applications are linear and much theory devoted to the "improvement" of linear estimates of linear discriminants. We also note that even for the particular normal distribution setup discussed here there has as yet not been any completely frequentist rule that guarantees optimal allocation when the parameters are unknown nor a Bayesian rule which yields V^* for all values of q_1, q_2, N_1 and N_2. On the other hand, when allocation is actually not the goal, linearity may be inherently more useful (certainly descriptively) because of its simplicity in discussing certain issues, and in the normal case both frequentist and Bayesian estimation procedures will yield linear sample discriminants.

4. Using Linear Discriminants--Normal Case.

As in the previous section let X_i, $i = 1, 2$ represent a set of N_i observations known to be from π_i, a $N(\mu_i, \Sigma)$ population. The object is to optimally allocate a new observation u which has prior probability q_i of being from π_i. We then assign a prior probability $g(\mu_1, \mu_2, \Sigma)$ to the unknown set of parameters. Hence

$$R = \frac{Pr[\pi_1|u]}{Pr[\pi_2|u]} = \frac{q_1 f(u|X, \pi_1]}{q_2 f(u|X, \pi_2]} \tag{4.1}$$

where $X = (X_1, X_2)$ and

$$f(u|X, \pi_i) = \int f(u|\mu_i, \Sigma) p(\mu_1, \mu_2, \Sigma|X) d\mu_1 d\mu_2 d\Sigma, \tag{4.2}$$

the predictive density of a future observation where

$$p(\mu_1, \mu_2, \Sigma|X) \propto L(\mu_1, \mu_2, \Sigma|X)\, g(\mu_1, \mu_2, \Sigma). \tag{4.3}$$

This then provides the solution for the allocation of the next observation and can be used on all further observations. This latter use is not optimal as the predictive distribution of a set of new observations is dependent and here it would be utilized as if they were independent (for the optimal solution see Geisser (1966)). At any rate the solution is optimal for the next observation u. However one is in quandary as how to calculate a joint prior distribution for μ_1, μ_2, and Σ that realistically reflects prior knowledge one may have about them. One way out of this dilemma is to

use the improper prior $g(\mu_1, \mu_2, \Sigma^{-1}) \propto |\Sigma|^{\frac{p+1}{2}}$ which tends to minimize the effect of the prior distribution. The results for this case were given by Geisser (1964) and yields for the posterior probability ratio

$$R = \frac{q_1}{q_2}\left(\frac{N_1}{N_2}\right)^{\frac{p}{2}}\left(\frac{N_1+1}{N_2+1}\right)^{\frac{N_1+N_2-1-p}{2}}\left[\frac{(N_2+1)(N_1+N_2-2)+N_2(u-\bar{x}_2)'S^{-1}(u-\bar{x}_2)}{(N_1+1)(N_1+N_2-2)+N_1(u-\bar{x}_1)^1S^{-1}(u-\bar{x}_1)}\right]^{\frac{N_1+N_2-1}{2}} \quad (4.4)$$

so that when $R > 1$ assign u to π_1 and to π_2 otherwise. This rule is in general quadratic and is linear only for very special cases among which is $N_1=N_2$ and $q_1=q_2$, but tends to linearity as the sample sizes increase.

All evidence to date indicates that this procedure is superior for allocation than the plug-in rule V^* of (3.6). In this regard admissibility was previously discussed. From the point of view of density estimation the predictive density, as generally suggested in Geisser (1971), is shown by Aitchison (1975) to be a better estimate of the true density in this case than what results from plugging in the maximum likelihood estimates into the known normal density (which is basically the rule V^*) by a "frequentist" goodness of fit criterion based on the Kullback-Leibler (1951) directed measure of divergence.

On the other hand the use of V (V^* with $q_1 = q_2$) as a separatory function can be made compelling or approximately so even when based on probabilistic criterion of the kind discussed in (2.10), Enis and Geisser (1974). Neither in its form nor its interpretation, is (4.4) very appealing for separatory purposes, while using V as a separatory function seems to be very attractive for many applications.

If one then were satisfied with V as a separatory function and decided to use V* as well in the allocatory mode as in most applications, what can we say about its properties, i.e., how good an allocator is it. Before answering this let us examine the allocatory prowess of U* when the parameters are known--the best possible situation. Then, letting θ stand for μ_1, μ_2 and Σ,

$$\text{PCC} = \gamma(\theta) = q_1\gamma_1(\theta) + q_2\gamma(\theta)$$

where

$$\left.\begin{aligned} \gamma_i(\theta) &= \Pr[U^* > 0 | \pi_1, \theta] \\ \gamma_L(\theta) &= \Pr[U^* < 0 | \pi_2, \theta], \end{aligned}\right\} \tag{4.5}$$

$\gamma_i(\theta)$ being the probability of U* correctly classifying an observation emanating from π_i. It is easily shown, Geisser (1967), that

$$\left.\begin{aligned} \gamma_1(\theta) &= 1 - \Phi(\tau_1) \\ \gamma_2(\theta) &= \Phi(\tau_2) \end{aligned}\right\} \tag{4.6}$$

where $\tau_1 = (\log q_2 q_1^{-1} - \frac{1}{2}\alpha)/\alpha^{\frac{1}{2}}$, $\tau_2 = (\log q_2 q_1^{-1} + \frac{1}{2}\alpha)/\alpha^{\frac{1}{2}}$ and $\alpha = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$.

Hence a "plug-in" estimate of the best one can do is $\gamma(\hat{\theta}) = q_1(1 - \Phi(\hat{\tau}_1)) + q_2\Phi(\hat{\tau}_2)$ and $\hat{\tau}_1$ and $\hat{\tau}_2$ are estimated by employing $Q = (\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2)$ as an estimate for α where μ_1, μ_2 and Σ are unknown. For a Bayesian estimate of $\gamma(\theta)$ which employs $E_\theta\gamma(\theta) = \bar{\gamma}$, see Geisser (1967, 1970). It must be noted that this is an

estimate of the best that can be done in terms of $\gamma(\theta)$ and not an estimate of what may be achieved with a given V* when the parameters are unknown. When one actually uses V* then we have the conditional or Actual PCC(APCC)

$$\text{APCC} = \delta(\hat{\theta}, \theta) = q_1\delta_1(\hat{\theta}, \theta) + q_2\delta_2(\hat{\theta}, \theta) \tag{4.7}$$

where

$$\left.\begin{aligned} \delta_1(\hat{\theta}, \theta) &= \Pr(V^* > 0 | \pi_1, \hat{\theta}, \theta) = 1 - \Phi(\eta_1) \\ \delta_2(\hat{\theta}, \theta) &= \Pr[V^* < 0 | \pi_2, \hat{\theta}, \theta) = \Phi(\eta_2) \end{aligned}\right\} \tag{4.8}$$

where

$$\eta_i = \{\tfrac{1}{2}(\bar{x}_1 + \bar{x}_2 - \mu_i)S^{-1}(\bar{x}_1 - \bar{x}_2) + \log q_2 q_1^{-1}\} \Big/ \left[(\bar{x} - \bar{x}_2)' S^{-1} \Sigma S^{-1} (\bar{x}_1 - \bar{x}_2)\right]^{\frac{1}{2}} \tag{4.9}$$

$i = 1, 2$. A naive estimate $\hat{\delta}(\hat{\theta}, \theta) = \delta(\hat{\theta}, \hat{\theta})$ turns out to be $\gamma(\hat{\theta})$, so that the estimate for APCC is the same as for the PCC which is most unsatisfactory and has led to much effort by frequentists in attempting to correct that estimate of APCC or its peculiar companion $E[\text{APCC}] = E_{\hat{\theta}}[\delta(\hat{\theta}, \theta)]$, Hills (1966). In fact for a long time the various possible probabilities of correct classification were confused and the subject in somewhat of a chaotic state until Hills (1966) presented a careful analysis of the various frequentist allocation error rates. For some further remarks see Geisser (1969, 1970). From a Bayesian point of view the problem as such completely disappears by using as estimators for $\delta_i(\hat{\theta}, \theta)$ its posterior expectation $\bar{\delta}_i = E_\theta(\delta_i(\hat{\theta}, \theta))$ which yields Geisser (1967)

$$\bar{\delta}_1 = \Pr\left[t_{N_1 + N_2 - 1 - p} > \frac{(\log q_2 q_1^{-1} - \frac{1}{2}Q)(N_1 + N_2 - 1 - p)^{\frac{1}{2}} N_1^{\frac{1}{2}}}{[(N_1 + N_2 - 2)(N_1 + 1)Q]^{\frac{1}{2}}}\right]$$

$$\bar{\delta}_2 = \Pr\left[t_{N_1 + N_2 - 1 - p} < \frac{(\log q_2 q_1^{-1} + \frac{1}{2}Q)(N_1 + N_2 - 1 - p)^{\frac{1}{2}} N_2^{\frac{1}{2}}}{[(N_1 + N_2 - 2)(N_2 + 1)Q]^{\frac{1}{2}}}\right] \quad (4.10)$$

where $t_{N_1 + N_2 - 1 - p}$ is the student "t" random variable with $N_1 + N_2 - 1 - p$ degrees of freedom. Clearly then $\bar{\delta} < \bar{\gamma}$ since $\delta(\theta) < \gamma(\theta)$ as required. Actually the Bayesian estimate of $\gamma(\theta)$ is rather difficult to compute explicitly but $\bar{\gamma} \doteq \gamma(\hat{\theta})$, for a better approximation see Geisser (1970). Note also that $\bar{\delta} < \gamma(\hat{\theta})$. At any rate a clear interpretation emerges—$\bar{\gamma}$ or $\gamma(\hat{\theta})$ is an estimate of what potentially could be achieved with sample sizes very much larger than those in hand whilst $\bar{\delta}$ is what is actually achievable with the data in hand. In other words if say $\gamma(\hat{\theta})$ is large enough, then the discriminatory variables are satisfactory. However the user may not be satisfied with an appreciably lower value of $\bar{\delta}$. But then the remedy is clear, one needs larger sample sizes until $\bar{\delta}$ is close enough to $\gamma(\hat{\theta})$ to be satisfactory. If $\gamma(\hat{\theta})$ is not large enough to suit the purposes of the allocation then one must find other discriminatory variables.

While $\bar{\delta} = q_1\bar{\delta}_1 + q_2\bar{\delta}_2$ is an estimate of the APCC, it turns out that $\bar{\delta}$ is exactly the predictive probability of correct classification (PPCC) if $N_1 = N_2$. When $N_1 \neq N_2$, $\bar{\delta}$ is approximately the PPCC for large N_1 and N_2, Geisser (1967). Further $\bar{\delta}$ is also a useful guide in determining which variables may be omitted in measuring future observations. For example it can happen for economic or other reasons that

only a subset of r of the p original variables can be utilized for allocating future observations. Then one could compute $\overline{\delta}$ weighted by an appropriate cost or utility factor for each subset of r out of the p variables in order to make an optimal determination.

At any rate this approach yields sensible answers when one uses the usual linear discriminant for allocation. Slight improvements can be made by some adjustment of V within the Bayesian framework as noted by Geisser (1967) and Enis and Geisser (1974), but it's not likely the effect will be significant. Extension to $r > 2$ populations throughout or q_i unknown presents no intrinsic difficulty, see Geisser (1964, 1967).

5. Maximizing Measures of Spread for Linear Discriminatory Forms.

When there are no appropriate distributional assumptions one can proceed by both choosing a class $\mathcal{D}(u)$ of discriminatory functions, linear, quadratic, etc. and defining either a distance between any two populations, Fisher [1936], or a more general measure of spread amongst all of the populations. A minimal set of "best" discriminants then presumably would be selected from all of those solutions that maximize the spread with respect to the parameters of the discriminatory functions given the constraints under consideration. These discriminants then can be used to completely characterize the differential aspects of the populations with respect to the manifest variables.

Let us further assume all of the distributions of the r populations are roughly the same in that they enjoy approximately the type of clustering and symmetry about their mean vectors exhibited by a set of multivariate normal densities with equal covariance matrices. Basically then the important differences are in the location of these central vectors. Fisher (1936) then found it sensible for r populations, to find the set of linear combinations $c'u$ which maximized pairwise the distance functions $\{c'(\mu_i-\mu_j)\}^2/c'\Sigma c$, $i\neq j=1, \ldots, r$ where Σ was assumed to be the common covariance matrix. This generates the optimal reduced set of linear discriminants previously obtained where multivariate normal theory was assumed. The technique used by Fisher was essentially differentiation with Lagrange multipliers. An alternate geometric derivation is given by Dempster [1969].

There are other methods of obtaining these linear discriminants, which involve maximizing some measure of spread, Wilks (1962). The technique used

for the maximization by Wilks also involved Lagrange multipliers. We now present an alternate derivation, Geisser (1973b) which is completely algebraic and somewhat more general. Again, suppose there are r p-dimensional multivariate populations with means $\mu_1, \ldots \mu_r$ and common positive definite covariance matrix Σ. Further let β, of rank $r-v < p$, be defined as previously in Section 2. Assume that $g(\beta\Sigma^{-1}) = g(\delta_1,\ldots,\delta_{r-v})$ is any scalar measure of the spread of these r populations that is increasing in the non-zero roots of $\beta\Sigma^{-1}$, $\delta_1 \geq \ldots \geq \delta_{r-v} > 0$. Suppose further we transform these r p-dimensional populations into a $k \leq p$ space by a real transformation matrix $C_{k\times p}$ which is of rank k. Hence $\eta = C\mu_i$, $i=1,\ldots,r$, $\Omega = C\Sigma C'$; $\Gamma = C\beta C'$ and the measure of spread in k dimensions is $g_k(\Gamma\Omega^{-1})$, i.e., the same scalar function of the non-zero roots $d_1 \geq \ldots \geq d_t > 0$ of $\Gamma\Omega^{-1}$ where $t = \min(k, r-v)$. Then we shall show that

$$\underset{C}{\text{Max}}\ g_k(\Omega^{-1}\Gamma) = g(\delta_1,\ldots,\delta_t)\ . \tag{5.1}$$

As the maximum spread is attained for $k = r-v$, there is no interest in the discriminatory situation in considering $k > r-v$. An orthogonal basis solution for C, when $k \leq r-v$, would then be

$$C = P'_{(k)}\Lambda\Sigma^{-1} \tag{5.2}$$

where $P_{(k)}$ is as previously defined. Consequently ΛP_j is the characteristic vector associated with the j^{th} largest root of $\beta\Sigma^{-1}$. Hence the conjecture made previously in Section 2 has a basis in fact if optimization depends on maximizing every scalar measure of spread which is an increasing function of the non-zero roots of $\beta\Sigma^{-1}$. One can also define the fraction of total loss sustained in the measure of spread when $k \leq r - v$ as

$$L = \frac{g(\delta_1,\ldots,\delta_{r-v}) - g(\delta_1,\ldots,\delta_k)}{g(\delta_1,\ldots,\delta_{r-v})} \tag{5.3}$$

For example, if we are using either the "Hotelling" or "Wilks" measure of spread:

$$g_H = \mathrm{Tr}\Sigma^{-1}\beta = \sum_{i=1}^{r-v}\delta_i \ , \ g_W = \left|I + \beta\Sigma^{-1}\right| = \prod_{i=1}^{r-v}(1+\delta_i) \tag{5.4}$$

then

$$L_H = \sum_{i=k+1}^{r-v}\delta_i \Big/ \sum_{i=1}^{r-v}\delta_i \ , \ L_W = 1 - \prod_{i=k+1}^{r-v}(1+\delta_i)^{-1} \tag{5.5}$$

The algebraic derivation of the aforementioned results is basically an application of the following matrix theorem.

Theorem: Let Z be a real $p \times m$ matrix of rank $s = \min(p,m)$ and E_k be the class of $p \times p$ real symmetric idempotent matrices of rank k. Then for all $F \in E_k$ the maximum attainable values of the first t ordered roots a_i of $Z'FZ$ are α_i, $i=1,\ldots,t$, $t = \min(k,m)$, where the α_i's are the non-zero ordered roots of $Z'Z$. Further, the totality of solutions for F, where the maximum values of the roots are attained is given by

$$F_o = \begin{cases} Y_{(k)}D_k^{-1}Y_{(k)}' & \text{for } k \le m \\ Z(Z'Z)^{-1}Z' + G_{k-m} & \text{for } k \ge m \end{cases} \tag{5.6}$$

where $Y_{(k)} = (Y_1,\ldots,Y_k)$, represents the first k columns of $Y = ZP$, and P is the orthogonal matrix such that $P'Z'ZP = D_m$ where

$$D_j = \begin{pmatrix} \alpha_1 & 0 \cdots & 0 \\ 0 & \ddots & \vdots \\ \vdots & \ddots & 0 \\ 0 & \cdots\ 0 & \alpha_j \end{pmatrix} \quad ;$$

and G_{k-m} is any idempotent matrix of rank k-m orthogonal to Z.

Proof:

Given the definitions of P and Y above then

$$\begin{aligned} D_m &= P'Z'ZP = P'Z'FZP + P'Z'(I-F)ZP \\ D_m &= Y'Y = Y'FY + Y'(I-F)Y. \end{aligned} \tag{5.7}$$

Hence the roots of $Z'FZ$ are the roots of $Y'FY$ and by virtue of (5.7) the ordered roots of $Y'FY$, $a_1 \geq \cdots \geq a_m \geq 0$, are not greater than the ordered roots of $Y'Y$, i.e., $\alpha_i \geq a_i$, $i = 1, \ldots, m$, Bellman [1960, p. 113].

Now let Q be a $p \times p$ orthogonal matrix such that $Q = (Y_{(s)}D_s^{-\frac{1}{2}}, Q_2)$, where $Y_{(s)}$ consists of the first s columns of Y. Then

$$Y'FY = Y'QQ'FQQ'Y = Y'QF^*Q'Y$$

where F^* is obviously idempotent since F is assumed to be. Further

$$Y'Q = \begin{array}{c} \\ s \\ m-s \end{array} \begin{array}{c} \begin{array}{cc} s & p-s \end{array} \\ \left(\begin{array}{c|c} D_s^{\frac{1}{2}} & 0 \\ \hline 0 & 0 \end{array}\right) \end{array} \tag{5.8}$$

so that

$$Y'FY = Y'QF^*Q'Y = \begin{array}{c} \\ s \\ m-s \end{array} \begin{array}{c} \begin{array}{cc} s & m-s \end{array} \\ \left(\begin{array}{c|c} D_s^{\frac{1}{2}}F_{11}^*D_s^{\frac{1}{2}} & 0 \\ \hline 0 & 0 \end{array}\right) \end{array} \tag{5.9}$$

and F_{11}^{*} is the $s \times s$ matrix in the upper left hand corner of F^{*}. Now the maximum rank of $Y'FY$ is $t = \min(k,m)$ or $t = \min(k,s)$. Hence all solutions, for which $a_i = \alpha_i$, $i=1, \ldots, t$ are such that the rank of $Y'FY$ must be t. Further the first t diagonal elements of F_{11}^{*} are then 1 since $\sum_{i=1}^{t}\alpha_i = \sum_{i=1}^{s}\alpha_i f_{ii}^{*}$, $0 \leq f_{ii}^{*} \leq 1$ and $\sum_{i=1}^{p} f_{ii}^{*} = k$ must be satisfied. This implies that the off diagonal elements in those rows and columns are zero since we are dealing with idempotent matrices. Therefore, all solutions for F^{*} are

$$F_o^{*} = \begin{array}{c} \\ k \\ p-k \end{array} \begin{array}{c} \begin{array}{cc} k & p-k \end{array} \\ \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right) \end{array} \qquad \text{for } t=k, \text{ i.e., } k \leq m$$

or (5.10)

$$F_o^{*} = \begin{array}{c} \\ m \\ p-m \end{array} \begin{array}{c} \begin{array}{cc} m & p-m \end{array} \\ \left(\begin{array}{c|c} I_m & 0 \\ \hline 0 & G \end{array}\right) \end{array} \qquad \text{for } t=m, \text{ i.e., } m \leq k,$$

where G is any idempotent $p-m \times p-m$ matrix of rank $k-m$. Hence the totality of solutions for F are $F_o = QF_o^{*}Q'$, so that

$$F_o = Y_{(k)}D_k^{-1}Y_{(k)}' \qquad \text{for } k \leq m \qquad (5.11)$$

which is unique if $\alpha_1, \ldots, \alpha_k$ are distinct and

$$F_o = YD_m^{-1}Y' + Q_2GQ_2' \qquad \text{for } m \leq k. \qquad (5.12)$$

Note that from (5.12) and $ZP = Y$ that

$$F_o = ZPD_m^{-1}P'Z' + Q_2GQ_2' = Z(Z'Z)^{-1}Z' + Q_2GQ_2' \quad \text{for} \quad m \leq k \ . \tag{5.13}$$

Hence set $Q_2GQ_2' = G_{k-m}$. Since G is an arbitrary idempotent matrix of rank $k-m$ and Q_2 is orthogonal to Z being it is orthogonal to Y , then G_{k-m} is an arbitrary idempotent matrix orthogonal to Z and the theorem is established.

As an immediate consequence of the theorem and the fact that $a_i \leq \alpha_i$ we have the following:

Corollary

If $g(Z'FZ) = g(a_1,\ldots,a_t)$ is a scalar non-decreasing function of the roots a_i , then

$$\max_{F \in E_k} g(a_1,\ldots,a_t) = g(\alpha_1,\ldots,\alpha_t) \ . \tag{5.14}$$

In order to apply the theorem and corollary we first note that the non-zero roots of $\Gamma\Omega^{-1} = C\beta C'(C \Sigma C')^{-1}$ are the same as the non-zero roots of $\Lambda'C'(C \Sigma C')^{-1}C\Lambda$ where $\beta = \Lambda\Lambda'$ and Λ is $p \times r\text{-}v$. Set $C \Sigma^{\frac{1}{2}} = H$ where $\Sigma^{\frac{1}{2}}$ is the positive definite symmetric square root of Σ so that the non-zero roots of $\Gamma\Omega^{-1}$ are the same as the non-zero roots of $\Lambda'\Sigma^{-\frac{1}{2}}H'(HH')^{-1}H\Sigma^{-\frac{1}{2}}\Lambda$. Set $r\text{-}v = m$, $Z = \Sigma^{-\frac{1}{2}}$ and the idempotent matrix $H'(HH')^{-1}H = F$. Hence as by our previous corollary

$$\max_C g_k(\Gamma\Omega^{-1}) = \max_F g_k(Z'FZ) = g(\delta_1,\ldots,\delta_t) \ .$$

To find solutions for C we note that there is an orthogonal matrix P such that

$$P'\Lambda'\Sigma^{-1}\Lambda P = \Delta_m = \begin{pmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_m \end{pmatrix}$$

and in the theorem set $Y = \Sigma^{-\frac{1}{2}}\Lambda P$ and $Y_{(j)} = \Sigma^{-\frac{1}{2}}\Lambda P_{(j)}$ where $P_{(j)} = (P_1, \ldots, P_j)$ is the matrix consisting of the first j columns of P. Note also that ΛP_i is the invariant vector of $\beta\Sigma^{-1}$ corresponding to the root δ_i, $i = 1, \ldots, m$ and the calculation of $Y_{(j)}$ does not depend on Λ. Hence from the theorem

$$F_o = Y_{(k)}\Delta_k^{-1}Y'_{(k)} \qquad \text{for } k \leq r - v. \tag{5.15}$$

From $H'(HH')^{-1}H = F$ we obtain $H = HF$ and noting from (5.15) that $Y'_{(k)}F_o = Y'_{(k)}$ then $H_o = Y'_{(k)}$ and $C_o = Y'_{(k)}\Sigma^{-\frac{1}{2}} = P'_{(k)}\Lambda'\Sigma^{-1}$ for $k \leq r - v$, as required.

The derivations in this section and in Section 2 have been presented in terms of population parameters. But obviously if sample estimates based on data at hand, $\hat{\beta}$ and $\hat{\Sigma}$ are utilized there need not be, from one point of view, any essential change other than "optimization" now takes place with regard to sample estimates. The problem then is to decide on the "plug-in" estimators. The substitution of unbiased sample moments for the population moments results in the same set of sample discriminants as is used in the normal case where maximum likelihood estimates corrected for bias are utilized. Of course the simplifying assumption that the multivariate populations differ mainly in their locations and relative to this, variations in the dispersion matrices were unimportant as exemplified explicitly in the previously discussed multivariate normal model, was our guide. This latter model gave rise to linear theory in terms of population discriminants and consequently, it is no surprise that focusing on linear theory can yield the same results.

6. Sample Reuse Techniques

When precise distributional assumptions are untenable or abandoned entirely so that theoretical calculations are precluded, one is then obliged to provide some other means for rendering discriminants and assessing their quality. In this section we shall discuss how sample reuse notions can be directed towards an evaluation of discriminants and their further refinement. Again for simplicity let us focus on the two population case. Let the usual single linear discriminant be $V = V(u)$. By substituting each of the p-dimensional observations x_{ij} for $j = 1, \ldots, N_i$ and $i = 1, 2,$ in V we obtain $V(x_{ij}) = v_{ij}$ or two sets of univariate observations. These may now be plotted on a single axis distinguishing them only by their population origin. If there are a great many of them a histogram for each set is visually informative in indicating the quality the linear separation induced. If V^* is to be used for allocatory purposes then some assessment of the APCC is in order. It has long been clear that the naive assessment of the APCC, by merely calculating $q_1 n_1 N_1^{-1} + q_2 n_2 N_2^{-1}$ where n_i represents the number of x_{ij}'s that are correctly classified by V^*, will be too large - just as in the normal case $\gamma(\hat{\theta})$ was generally too large as discussed in section 4. A sample reuse technique for correcting this flaw was proposed by Lachenbruch (1965). He proposed calculating $V^*_{ij}(u)$ the linear discriminant with x_{ij} omitted and then computing $V^*_{ij}(x_{ij}) = u_{ij}$ and classifying x_{ij} on the basis of whether u_{ij} exceeded or fell short of 0. An adjusted estimate of APCC, $q_1 n_1' N_1^{-1} + q_2 n_2' N_2^{-1}$, is obtained where n_i' represents the number of x_{ij}'s, $i = 1, 2,$ correctly classified. Note that if $N_i/(N_1 + N_2)$ is appropriate as an estimator of q_i,

when it is unknown, that the estimator for the APCC becomes $(n_1' + n_2')/(N_1 + N_2)$. This is reasonable in situations where the initial sample of size $N = N_1 + N_2$ is drawn at random from $\Pi = \Pi_1 \cup \Pi_2$, so that the random frequency N_i provides information on q_i.

This method may also be of value in the determination of which variables could be eliminated or which subset of r out of the p would be optimal in future measurements as discussed in section 4.

One could attempt a finer tuning, as it were, by applying the predictive sample reuse (PSR) method, Geisser (1975). A criterion applicable here would be to maximize

$$P(u_0) = q_1 n_1(u_0) N_1^{-1} + q_2 n_2'(u_0) N_2^{-1} \tag{6.1}$$

with respect to u_0, where $n_i(u_0)$ represents the number of x_{ij}'s correctly allocated by V_{ij}^* such that x_{ij} is allocated to Π_1 if $V_{ij}^*(x_{ij}) = u_{ij} > u_0$ and to Π_2 otherwise. One would order the scalar values u_{ij} and find that cutoff point $\hat{u}_0$ which maximizes $P(u_0)$. This can easily be done numerically as it is essentially a counting procedure. Convenient algorithms can be found to shorten the process. While it is also clear that $\hat{u}_0$ need not be unique, it can be made so by arbitrarily selecting a particular one of them, e.g. the maximizer $\hat{u}_0$ closest to zero. Then for future allocation one uses $V^*(u) \gtrless \hat{u}_0$ as the allocatory discriminant. One could also alter the criteria when q_i is unknown and maximize $\hat{P}(u_0) = \hat{q}_1 n_1(u_0) N_1^{-1} + \hat{q}_2 n_2(u_0) N_2^{-1}$. If $N_i(N_1 + N_2)^{-1}$ can be used as an estimator for q_i then the new criterion effectively maximizes the total number of x_{ij}'s correctly classified by $\hat{V}_{ij}^* \gtrless u_0$, and again one would use $\hat{V}^*(u) \gtrless \hat{u}_0$ as the allocator where $\hat{V}_{ij}^*$ and $\hat{V}^*(u)$ are merely V_{ij}^* and $V^*(u)$ respectively with $q_1 q_2^{-1}$ replaced by $N_1 N_2^{-1}$.

An appropriate assessment of the new discriminant would require a two-deep cross-validatory assessment i.e. of the Lachenbruch (1965) type. There are obviously ways of applying the PSR approach to linear discriminants other than the cut-off point allocatory approach illustrated here, e.g. estimating by PSR the linear regression coefficients.

If one is concerned mainly with the reliability of a discriminant in its separatory role, then Mosteller and Tukey (1968) and Lachenbruch and Mickey (1968) suggest jackknifing V. First one calculates the set of pseudo-discriminant functions $V'_{ij} = (N_1 + N_2)V - (N_1 + N_2)V_{ij}$ $j = 1, \ldots, N_i$, $i = 1, 2$, and then $V' = (N_1 + N_2)^{-1} \sum_{i,j} V'_{ij}$, which is termed the jackknifed discriminant. One can compute the reliability of V' in terms of the variation of the regression coefficients of the V'_{ij}, the individual values averaged to compute V'. Examining the ratio of a regression coefficient in V' to its sample standard error permits a judgement on the significance of its deviation from zero. The main point of this exercise is to assess to some degree the reliability of the jacknifed discriminant function in its separatory role. Again if one decides to use V' (or $V^{*\prime}$) as an allocator one can assess it by using a two-deep cross-validatory approach.

7. Remarks.

I have attempted to carefully delineate two distinct purposes of discriminatory analyses and to examine the linear aspects involved. However linearity has been surveyed, as it were, from a personal (not to be confused with personalistic) point of view, in that much work on the linear aspects of normal parametric discrimination has not been mentioned chiefly because it involves the generation of modest improvements with regard to certain frequency properties by some slight alteration of the linear discriminants. It is my contention that here the Bayesian approach, or when adjustments are indicated, a Bayesian type of adjustment will yield better results than frequential tinkering. When parametric assumptions are fuzzy or nonexistent, sample reuse methods, which are frequency oriented predictive simulation techniques, should serve.

Finally it must be borne in mind that Discrimination is a technique which is often most useful in the early history or soft stage of a discipline when notions are fuzzy, measurements crude or indirect and relationships vaguely understood at best. Hence it is generally an appreciable improvement of whatever has gone before--theoretical niceties notwithstanding. No doubt Linear Discrimination fulfills the role played by Barnard's (1972) "midwife" in fostering the parturition of pertinent distinctions, probabilistic or classificatory, during the birthpangs of a scientific discipline--but soon abandoned or its focus shifted as the discipline hardens.

References

Aitchison, J. (1975). Goodness of prediction fit. Biometrika, 62, 3, pp. 547-554.

Anderson, J. A. (1973). Logistic discrimination with medical applications. Discriminant Analysis and Applications. edited by T. Cacoullos. New York: Academic Press. pp. 1-16.

Anderson T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons.

Barnard, G. A. (1972). The unity of statistics. J. R. Statist. Soc. A., 135, pp. 1-14.

Bellman R. (1960). Introduction to Matrix Analysis, New York: McGraw-Hill.

Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations. Ann. Math. Statist., 36, pp. 1174-1184.

Dempster, A. P. (1969). Elements of Continuous Multivariate Analysis. Reading, Massachusetts: Addison-Wesley.

Desu, M. M. and Geisser, S. (1973). Methods and applications of equal-mean discrimination. Discriminant Analysis and Applications. edited by T. Cacoullos. New York: Academic Press. pp. 139-161.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. J. Am. Statist. Assoc. 70, 352, pp. 892-898.

Enis, P. and Geisser, S. (1970). Sample discriminants which minimize posterior squared error loss. South African Statist. J., 4, pp. 85-93.

Enis, P. and Geisser, S. (1974). Optimal predictive linear discriminants. Ann. Statist., 2, 2, pp. 403-410.

Fisher, R. A. (1936). The statistical utilization of multiple measurements. Annals of Eugenics, 8, pp. 311-335.

Geisser, S. (1964). Posterior odds for multivariate normal classification. J. R. Statist. Soc. B, 1,pp. 69-76.

———— (1966). Predictive discrimination. Multivariate Analysis. edited by P. Krishnaiah. New York: Academic Press. pp. 149-163.

———— (1967). Estimation associated with linear discriminants. Ann. Math. Statist. 38, pp. 807-817.

———————— (1969). Alternative views of discrimination. Bull. Int. Inst. Statist., 2, pp. 132-134.

———————— (1970). Discriminatory practices. Bayesian Statistics, edited by D. Meyer and R. C. Collier, Illinois: Peacock, pp. 57-70.

———————— (1973a). Multiple birth discrimination. Multivariate Statistical Inference. edited by D. G. Kabe and R. P. Gupta. Amsterdam:North-Holland, pp. 49-55.

———————— (1973b). A note on linear discriminants. Bull. Int. Statist. Inst. 1, pp. 442-448.

———————— (1975). The predictive sample reuse method with applications. J. Amer. Statist. Assoc., 70, 350, pp. 320-328.

Geisser, S. and Desu, M. M. (1968). Predictive zero-mean uniform discrimination. Biometrika, 55, 3, pp. 519-524.

Hills, M. (1966). Allocation rules and their error rates. J. R. Statist. Soc. B, 28, pp. 1-31.

Kiefer, J. and Schwartz, R. (1965). Admissible Bayes character of T^2, R^2 and other fully invariant tests for classical multivariate normal problems. Ann. Math. Statist. 36, pp. 747-770.

Kullback, S. and Liebler, R. A. (1951). On information and sufficiency. Ann. Math. Statist. 22, pp. 525-540.

Lachenbruch, P. A. (1965). Estimation of error rates in discriminant analysis. unpublished dissertation, University of California at Los Angeles.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics 10, pp. 1-11.

Mostetler, F. and Tukey, J. W. (1968). Data analysis, including statistics. Handbook of Social Psychology, edited by G. Lindzey and E. Aronson. Reading, Massachusetts: Addison-Wesley.

Okamoto, M. (1961). Discrimination for variance matrices. Osaka Math. J., 13, pp. 1-39.

Reyment, R. A. (1973). The discriminant function in systematic biology. Discriminant Analysis and Applications. edited by T. Cacoullos. New York: Academic Press. pp. 311-338.

Richter, D. L. and Geisser, S. (1960). A statistical model for testing treatment effects in the presence of learning. Biometrics, 16, 1, pp. 110-114.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statist. 15, 2, pp. 145-162.

Welch, B. L.(1939). Note on discriminant functions. Biometrika, 31, pp. 218-220.

Wilks, S. S. (1962). Mathematical Statistics. New York: John Wiley and Sons.